Unit Wise IMP Question bank

Chapter -1

1. Explain why data warehouses are needed for developing business solutions from today's perspective. Discuss the role of data marts. **(GTU- 2019, 2016)**
2. Compare data mart and data warehouse. **(GTU- 2019, 2017, 2015)**
3. Discuss Following: **(GTU- 2019, 2018, 2015)**
    a) Meta Data
    b) Virtual Warehouse
    c) Data Mart
    d) Enterprise Warehouse
    e) Business Intelligence
4. A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data – Justify. **(GTU- 2019, 2016, 2015)**
5. Can BI is used for DM? Or vice versa? Justify. **(GTU- 2019)**

Chapter -2

1. Discuss data warehouse architecture in detail. (GTU- 2019,2018, 2017)
2. Differentiate between OLTP and OLAP. Or Compare OLAP and OLTP in detail. **(GTU- 2019, 2017)**
3. Define data cube and explain 3 operations on it. **OR** What is the need of Online Analytical Processing (OLAP)? List categories of OLAP and explain any one. **OR** With the help of a suitable example, illustrate the OLAP operations: 'drilldown', 'roll-up', 'slice' and 'dice'. **(GTU- 2018, 2017)**
4. Differentiate between Operational Database System and Data Warehouse. **(GTU- 2018)**
5. Draw and Explain Snowflakes and Fact constellations Schema. Or Discuss star schema and fact constellation schema with diagram. **(GTU- 2019, 2018, 2017)**
6. Differentiate Fact table vs. Dimension table. **(GTU- 2019, 2018)**

7. What is apex cuboid? Discuss drill down and roll up operation with diagram. **(GTU- 2019)**

Chapter -3

1. Define data mining and list its features. **(GTU- 2019)**
2. What is Data Mining? Why is it called data mining rather knowledge mining? **(GTU- 2018, 2017)**
3. Describe the steps involved in data mining when viewed as a process of knowledge discovery. OR Explain KDD process with neat diagram. **(GTU- 2019, 2018, 2017, 2016)**
4. What is the difference between KDD and Data Mining? **(GTU- 2018)**
5. **Discuss possible ways for integration of a Data Mining system with a Database or Data Warehouse system.** (GTU- 2019)
6. Explain different types of Databases on which data mining can be performed. **(GTU- 2012, 2016, 2018)**
7. Explain the major issues in data mining. **(GTU- 2019, 2017, 2016, 2015)**

Chapter -4

1. Describe Concept Hierarchy? List and briefly explain types of Concept Hierarchy. **(GTU- 2019, 2018, 2017)**
2. In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem. **(GTU- 2019, 2018, 2017, 2016)**
3. What is noisy data? Explain the different techniques to remove the noise from data. **Or** How to handle nosy data? **Or** In data pre-processing why we need data smoothing? Discuss data smoothing by Binning. (GTU- 2019, 2018, 2017)
4. Enlist data reduction strategies and explain any two. (GTU- 2019, 2017)
5. Discuss attribute subset selection. Or Explain the following as attribute selection measure: (i) Information Gain (ii) Gain Ratio **(GTU- 2019, 2018, 2017)**

6. Explain Normalization Techniques with Example. **(2019, 2018)**
7. Explain Mean, Median, Mode, Variance, Standard Deviation & five number summary with suitable database example. **(GTU- 2019, 2018, 2017)**
8. Is Graphical visualization is better than text data? Justify your answer and explain different data visualization technique. **(GTU- 2019)**

Chapter -5

1. What is market basket analysis? Explain the two measures of rule interestingness: support and confidence. **(GTU- 2019, 2018, 2017)**
2. Consider the following dataset and find frequent item sets and generate association rules for them using Apriori Algorithm. minimum support count is 2 minimum confidence is 60%. **(GTU- 2019, 2018, 2017, 2016)**

TID	items
T1	I1, I2 , I5
T2	I2,I4
T3	I2,I3
T4	I1,I2,I4
T5	I1,I3
T6	I2,I3
T7	I1,I3
T8	I1,I2,I3,I5
T9	I1,I2,I3

3. Write and discuss the algorithm which is used to generate frequent itemsets using an iterative level-wise approach based on candidate generation. **(GTU- 2019, 2018, 2017)**
4. Discuss Hash-based technique to improve efficiency of Apriori algorithm. **OR** What are the limitations of Apriori algorithm and how can we increase the efficiency of Apriori Algorithm? **(GTU- 2019, 2017)**
5. Compare FP-Growth approach with Apriori.
6. Discuss the following terms: **(GTU- 2019, 2018, 2017)**
    a) Support
    b) Confidence
    c) Association Rules
7. What is Association Rule Mining? What are the types of Association

Rules? **(GTU- 2018, 2017)**

Chapter -6

1. Briefly explain Linear and Non-linear regression. **(GTU- 2019, 2018)**
2. What is classification and prediction? Explain issues regarding classification and prediction. **(GTU- 2018, 2017)**
3. Short note: Information gain, Gain Ratio, Gain Index. **(GTU- 2016, 2017)**
4. Briefly outline the major steps of decision tree classification. Why tree pruning useful in decision tree induction? **(GTU- 2019, 2018)**
5. Draw the topology of a multilayer, feed-forward Neural Network. Or Discuss various layers of multilayer feed-forward neural network with diagram. **(GTU- 2019, 2018)**
6. Why naïve Bayesian classification is called "naïve"? Describe naïve Bayesian classification with example. **(GTU- 2019, 2017)**
7. Explain various conflict resolution strategies in rule based classification. **(GTU- 2019)**
8. What is classification? Explain classification as a two step process with diagram. **(GTU- 2019, 2018)**
9. Discuss following terms. 1) Supervised learning 2) Correlation analysis 3) Tree pruning **(GTU- 2019)**

Chapter -7

1. Explain data mining application for fraud detection. Or Discuss fraud detection and click-stream analysis using data mining. **(GTU- 2019, 2018, 2017)**
2. Briefly explain the life-cycle of Data Analytics and discuss the role of data scientists. **(GTU- 2019, 2017)**

3. Explain role of Business intelligence in any one of following domain: Fraud Detection, Market Segmentation, retail industry, telecommunications industry. Explain how data mining can be helpful in any of these cases. **(GTU- 2017)**

Chapter -8

1. What is Big Data? What is big data analytic? Explain the big data distributed file system. **(GTU- 2019, 2018, 2017, 2016)**
2. Define Big Data. Discuss various applications of Big Data. **(GTU- 2019, 2018, 2017)**
3. Write a short note on Web mining. **(GTU- 2018, 2017)**
4. Explain Hadoop Architecture **(GTU- 2019, 2018, 2017)**
5. Explain Hadoop storage – HDFS. **(GTU- 2017)**
6. Explain text mining using example**. (GTU- 2019, 2017)**
7. What is meant by "clustering"? Explain why clustering is called unsupervised learning. Mention any two applications of clustering. (**GTU- 2019, 2017)**
8. Explain cluster analysis and outlier analysis with example. **(GTU- 2019, 2017)**
9. What do you mean by learning-by-observation? Explain k-Means clustering algorithm in detail. **(GTU- 2019, 2018)**

Twisted Question Bank

1. Minimum salary is 20,000Rs and Maximum salary is 1,70,000 Rs. Map the salary 1,00,000 Rs in new Range of (60,000 , 2,60,000) Rs using min-max normalization method.
2. If Mean salary is 54,000Rs and standard deviation is 16,000 Rs then find z score value of 73,600 Rs salary.
3. What is noise? Explain data smoothing methods as noise removal technique to divide given data into bins of size 3 by bin partition (equal frequency), by bin means, by bin medians and by bin boundaries. Consider the data: 10, 2, 19, 18, 20, 18, 25, 28, 22
4. Why naïve Bayesian classification is called "naïve"? Describe naïve Bayesian classification with example.
5. Suppose a group of sales price records has been sorted as follows: 6, 9, 12, 13, 15, 25, 50, 70, 72, 92, 204, 232 Partition them into three bins by equal-frequency (equi-depth) partitioning method. Perform data smoothing by bin mean.
6. In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.
7. With the help of a neat diagram explain the topology of a multilayer, feedforward Neural Network. Also explain the terms: "activation function" and "epoch".
8. Find all frequent item sets & generate strong association rules in the following database using Apriori Algorithm. Take minimum support count = 2 and min_conf = 70%

TID	Items
T100	I1, I2, I3,I4, I5
T200	I2, I3
T300	I1, I2, I6
T400	I2, I1, I7
T500	I1, I6, I8

9. Find All Frequent Item set using Apriori Algorithm. (Take minimum support count 2)

TID	ITEMS
T001	I1, I3, I4
T002	I2, I3, I5
T003	I1, I2, I3, I5
T004	I2, I5

10. Find all frequent itemsets in the following database using Apriori Algorithm. Take minimum support count = 2 and min_conf = 70%

TID	List of item_ids
T1	I1,I2,I3
T2	I1,I2
T3	I2,I5
T4	I1,I2,I5

11. For given following transaction database (Ref. Table 1), find out frequent itemsets using Apriori algorithm. Assume Minimum Support count=2 and Minimum Confidence=70%. Generate strong association rules from frequent itemsets.

General Question Bank

1. What is Cuboid? Explain any three OLAP Operations on Data Cube with example.
2. Compare OLTP & OLAP systems. **OR** Differentiate between OLTP and OLAP.
3. List out the different schema of data warehouse and explain one of the data warehouse schemas in detail with suitable diagram.
4. Draw and explain the data warehouse architecture.
5. Explain Star, Snowflake, and Fact Constellation Schema for Multidimensional Database.
6. What is Data Mining? Why is it called data mining rather knowledge mining?
7. Explain KDD process.
8. What is KDD? With the help of a neat diagram explain the steps in the KDD process.
9. Explain measures for finding rule interestingness (support, confidence) with example.
10. What is Data Mining? Write down short note on KDD process.
11. What are the major issues in Data Mining?
12. Explain what is meant by 'multi-dimensional data model' for data warehouse.
13. Discuss Star and Snowflake schemas using suitable examples.
14. Explain the need for Data Reduction during pre-processing and discuss various strategies for data reduction.
15. Compare association and classification. Briefly explain associative classification with suitable example.
16. Explain Mean, Median, Mode, Variance, Standard Deviation & five number summary with suitable database example.
17. What is Concept Hierarchy? List and explain types of Concept Hierarchy.
18. Draw and explain the data mining architecture.
19. What is data cleaning? Describe the different methods of handling missing values during data cleaning.
20. Explain the major issues in data mining.
21. What is data transformation? Explain the different data transformation approaches for transforming data.
22. Describe the various criteria for the classification of frequent pattern

mining. Explain the following as attribute selection measure: (i) Information Gain (ii) Gain Ratio

23. Write and explain the algorithm for mining frequent itemsets without candidate itemsets generation with suitable example.
24. Explain the various methods which evaluate the accuracy of a classifier or a predictor.
25. Explain the typical requirements of clustering in data mining.
26. Describe and explain the different types of clustering methods.
27. Write a short note on text mining.
28. List and describe methods for handling missing values in data cleaning.
29. What is the importance of visualization of discovered patterns? Explain the role of presentation in pattern visualization. Discuss various visualization techniques in KDD.
30. Define following terms & differentiate them: Data Mart , Enterprise Warehouse & Virtual Warehouse
31. Define the following terms: Business Intelligence, Data Mart, Closed frequent itemset, Outlier Analysis
32. Explain role of Business intelligence in any one of following domain: Fraud Detection, Market Segmentation, retail industry, and telecommunications industry. Explain how data mininig can be helpful in any of these cases.
33. Explain Prepruning and Postpruning with an example.
34. Explain the steps of the Apriori algorithm for mining frequent itemsets with candidate generation.
35. Explain various methods to improve the efficiency of the Apriori- based mining.
36. Write a short note on 'Linear Regression'.
37. Discuss density-based clustering in brief. Name any two algorithms for density based clustering.
38. Explain the working of the 'k-means' clustering algorithm. Also state its limitations.
39. Discuss the challenges in mining the World Wide Web, and explain the role of search engines.
40. What is outlier analysis? Why outlier mining is important? Briefly describe the different approaches: statistical-based outlier detection, distance-based outlier detection and deviation- based outlier detection.

41. Explain how trend analysis can be performed by mining Time-Series data.
42. Discuss the characteristics and limitations of Neural Networks. Mention any two suitable applications of neural networks.
43. What is Decision Tree? Explain how classification is done using decision tree induction.
44. What is Big Data? What is big data analytic? Explain the big data-distributed file system.
45. Define Big Data. Discuss various applications of Big Data.
46. Discuss applications of data mining in Banking and Finance.
47. Discuss the main features of Hadoop Distributed File System.
48. Explain Hadoop Architecture.
49. Write a short note on Web mining.